

A WEB-ACCESSIBLE FRAMEWORK FOR THE AUTOMATED STORAGE AND TEXTURE ANALYSIS OF BIOMEDICAL IMAGES

Michael Barnathan, Jingjing Zhang, Vasileios Megalooikonomou

Data Engineering Laboratory (DEnLab), Department of Computer and Information Sciences
Temple University, 1805 N. Broad St. Philadelphia, PA 19122

ABSTRACT

We present a framework for automated image texture analysis that utilizes Vector Quantization (VQ), an image compression technique, to perform common data mining operations, such as classification, clustering, and similarity searches, on 2D and 3D image datasets. We additionally demonstrate the effectiveness of this framework in a medical imaging context through MIDMS (Medical Image Data Mining System), a web-based system written in Perl and Matlab. MIDMS is capable of automating submission, normalization, compression, and real-time querying of general user-submitted medical image data. Our framework processes submissions by generating a locally optimal codebook for submitted datasets using the Generalized Lloyd Algorithm (GLA). After generating the codebook, our framework uses the codeword usage frequency of each image as the image's feature vector. Users of the framework may then perform highly accurate classification and similarity search experiments on these vectors using either the histogram model (HM) or summed Euclidean distance (SED) metrics, as confirmed by previous experiments utilizing these techniques. Our framework has the potential to assist researchers and clinicians in the sharing, mining, and analysis of large quantities of medical imaging data.

Index Terms— Texture descriptors, Vector quantization, Pattern analysis, Classification, Similarity Searches.

1. INTRODUCTION

Collecting and analyzing images acquired from many different sources presents a significant challenge within digital image processing and related fields, especially in a medical imaging context. Advancements in medical imaging modalities, such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT), have made large amounts of two-dimensional (2D) and three-dimensional (3D) image data available to researchers and clinicians. However, this abundance of data has necessitated advanced tools to organize, share, and analyze these images. In previous work,

we have analyzed feature extraction techniques based on image texture [1,2] and network topology of anatomical tree-like structures [3]. Previous results have confirmed the efficacy of these techniques, with classification accuracy reaching 96% in texture analysis of the breast ductal network and 89% in Functional Magnetic Resonance Imaging (fMRI) images of the brain. In this paper, we demonstrate how to construct a widely-accessible system around these techniques to facilitate the analysis, sharing, and storage of large amounts of medical imaging data. We specifically designed it for use on 2D and 3D images of the brain; however, our framework is modular and may be used for texture-based analysis of other types of medical images as well.

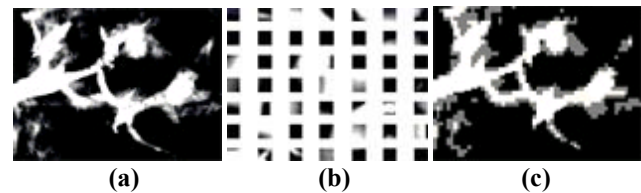


Fig. 1: (a) A breast ductal tree, (b) the codebook generated by GLA, and (c) the quantized image. Our approach yields high accuracies for images of the breast ductal network as well as brain images.

2. BACKGROUND

Our analysis utilizes techniques based on Vector Quantization (VQ) that we have developed previously [1,2,5,6]. The VQ approach is based on applying the keyblock image encoding [4] on submitted data to obtain compressed images. The keyblock approach decomposes each image into equi-size blocks and uses VQ to represent each block with the closest codeword from a codebook, as shown in Figure 1. First, given a fixed block size, each image is decomposed into a number of small blocks. Each small block contains features of the sub-area of its corresponding image. Based on such small blocks from

different images, a codebook containing keyblocks is generated. In order to generate the codebook, the Generalized Lloyd Algorithm (GLA), which produces a “local optimal” codebook based on the nearest neighbor and the centroid conditions, is used. The nearest neighbor condition formally states that two codewords x and y may only quantize to the same codeword if all nearer neighbors of x quantize to that codeword, where “nearness” is defined by a “distortion” function, typically Mean Squared Error. The centroid condition dictates that the generated codewords must be the centroids of their clusters. Starting with an initial codebook, GLA improves the codebook by iteratively applying the two conditions until the average distortion drops below a given threshold. Then, each image is encoded using the codebook. Initially, each image is decomposed into blocks, then for each block, the closest entry in the codebook is located and the corresponding index is stored. In such a way, each image is represented as a vector of frequencies of keyblock (codeword) appearance. Finally, the Histogram Model or Summed Euclidean Distance is employed as a similarity measure in classification and similarity searches.

3. ARCHITECTURE

We designed the user interface of our system in Perl due to the language’s powerful string manipulation capabilities, modularity, and ability to interface with web servers through CGI. We chose to make our system web-based to promote widespread access to our data, methods, and general framework and to encourage collaboration between researchers utilizing our system. The link between the user interface and the implementation of our methodology, which is written in Matlab, was established using the `Math::Matlab::Local` Perl module. This organization lends modularity to our framework, as our Matlab code may be decoupled from the interface and a different backend attached simply by changing the Matlab function that is called. As a result, our framework is suitable not only for the analysis of brain or breast images, the applications in mind when it was designed, but other types of images within the medical field. We used HTML template files with embedded Perl fragments, as processed by the `Text::Template` Perl module, to further decouple presentation and methodology, allowing the interface to be modified without changing the underlying analysis.

The complete source code to our system can be found at [7], but an example of how to join the two systems (without any error checking or input validation) is shown below:

```
my $matlab = new Math::Matlab::Local;
$matlab->cmd('/path/to/matlab');

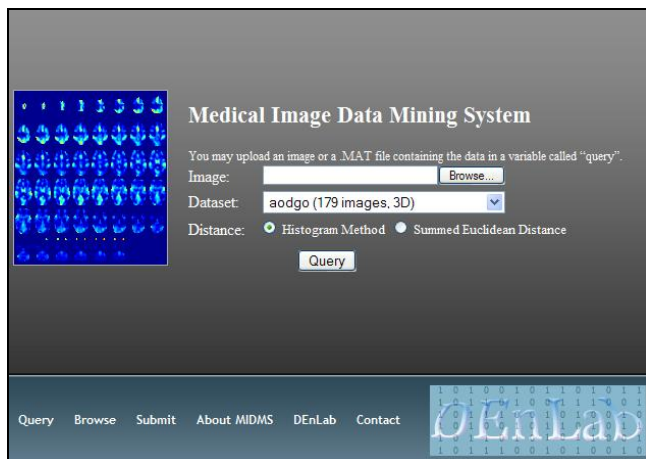
#Matlab requires a home directory, even
#if CGI does not run as a user.
$ENV{'HOME'} = '/tmp';

#Link to Matlab and load the dataset.
$matlab->execute("load
'datasets/$dataset/data.mat';
imwrite('images/$dataset.png')");

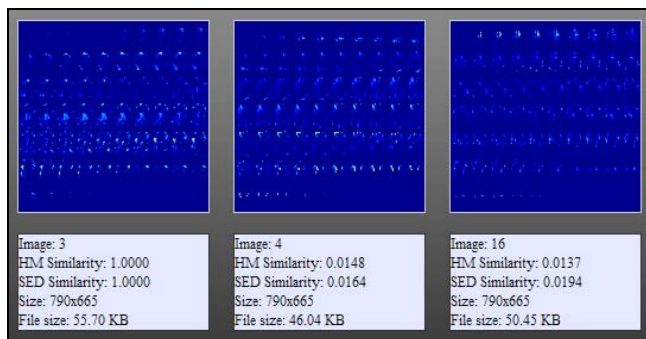
#The generated image is a regular file
#and can be accessed by any process with
#permissions, including Perl.
#For example:
open(DIMG, "<images/$dataset.png");
```

4. FRAMEWORK

When querying the dataset, users may upload a query image in any format readable by Matlab’s `imread()` function, including BMP, GIF, JPG, TIF, and PNG. Users may also select the dataset to query from a dropdown list of datasets and the number of images in each, and may choose between the Histogram Model and Summed Euclidean distance metrics. When the user clicks the “query” button, a similarity query is performed using the selected approach and thumbnails of the results from the dataset are displayed in descending order of similarity from the query using the specified metric. To balance processing time, storage, and transfer time, these thumbnails are automatically generated, resized, and cached in lossless 24-bit PNG format (as lossy formats, such as JPG, may induce artifacts that are unacceptable for inspection of medical images) upon submission of each dataset, while the original images are individually retrieved from the Matlab dataset when requested by a user clicking the corresponding thumbnail image. See Figure 2 for an example of the query process.



(a)



(b)

Fig. 2: (a) Querying a 3D dataset and (b) some results.

Users may also submit datasets to our system. When a user selects this option, he is prompted for a Matlab .MAT file containing his data as well as a name for the dataset, the number of codewords to generate in the codebook, the keyblock size, the number of sample images to reference in codebook generation, and a distortion threshold for termination of the Generalized Lloyd Algorithm. Codebook generation can be a lengthy process; therefore, when a user submits a dataset, it is not immediately analyzed, but rather moved to a temporary storage area, where it is processed by a script running periodically (as a cron job). Once the codebook has been created, the dataset is merged into the system and becomes available for querying.

5. CONCLUSION

Coupled with our previous techniques for analysis, our framework allows users to mine, share, and analyze large collections of medical images quickly and accurately. Our system's modularity additionally makes it suitable for a wide variety of uses in other biomedical and image processing applications. Opportunities for future work include integrating this platform with a two-step wavelet analysis technique we have previously developed, expanding the scope and generality of our system to allow researchers to easily perform complex, customized, and domain-specific experiments on submitted data, and further optimization of queries and codebook generation.

6. REFERENCES

- [1] Zhang, J., Megalooikonomou, V., An Effective and Efficient Technique for Searching for Similar Brain Activation Patterns, Proceedings of the IEEE International Symposium on Biomedical Imaging, Washington DC, Apr. 2007.
- [2] Megalooikonomou, V., Zhang, J., Kontos, D., Bakic, PR., Analysis of Texture Patterns in Medical Images with an Application to Breast Imaging, Proceedings of the SPIE Conference on Medical Imaging, San Diego, CA, Feb. 2007.
- [3] Megalooikonomou, V., Kontos D., Danglemaier J., Javadi A., Bakic P., Maidment A., A Representation and Classification Scheme for Tree-like Structures in Medical Images: An application on Branching Pattern Analysis of Ductal Trees in X-ray Galactograms, Proceedings of the SPIE Conference on Medical Imaging, San Diego, CA, Feb. 2006.
- [4] Mitra, S., Yang, S.Y., High Fidelity Adaptive Vector Quantization at Very Low Bit Rates for Progressive Transmission of Radiographic Images, Journal of Electronic Imaging, Vol. 8, pp.23-35, 1999.
- [5] Wang Q., Megalooikonomou V., Kontos D., A Medical Image Retrieval Framework, Proceedings of the 2005 IEEE International Workshop on Machine Learning for Signal Processing (MLSP05), Mystic, Connecticut, Sept. 28-30, 2005, pp. 233-238.
- [6] Kontos, D., Megalooikonomou, V., Fast and effective characterization for classification and similarity searches of 2D and 3D spatial region data, Pattern Recognition, Vol. 38, No. 11, pp. 1831-1846, 2005.
- [7] Barnathan, M., Zhang, J., Megalooikonomou, V. Medical Image Data Mining System. <<http://denlab.temple.edu/midms>>.